MBC 638
DATA ANALYSIS AND DECISION MAKING

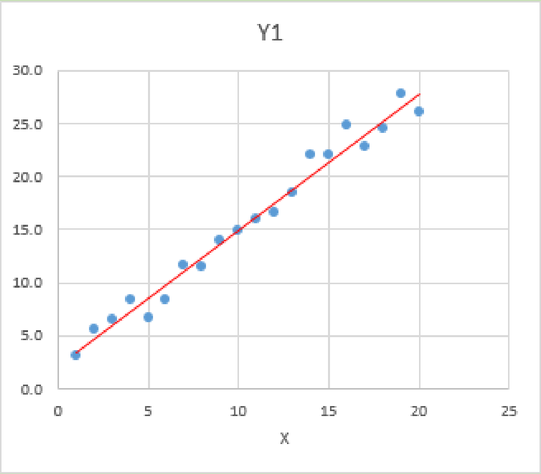
Anna Chernobai, Department of Finance

HYPOTHESIS TESTING in regressions
Chapters 10, 11

Hypothesis testing in regressions

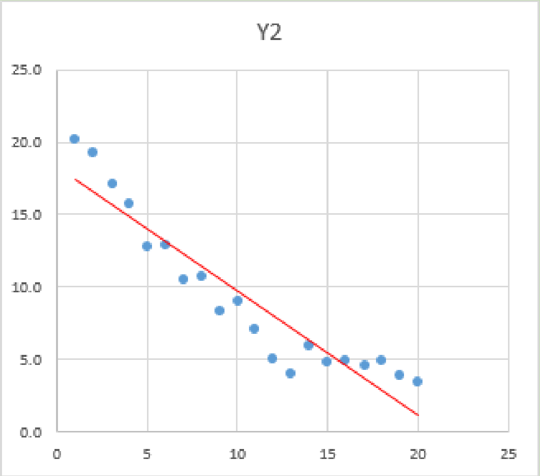
Linear relationship exists:

positive



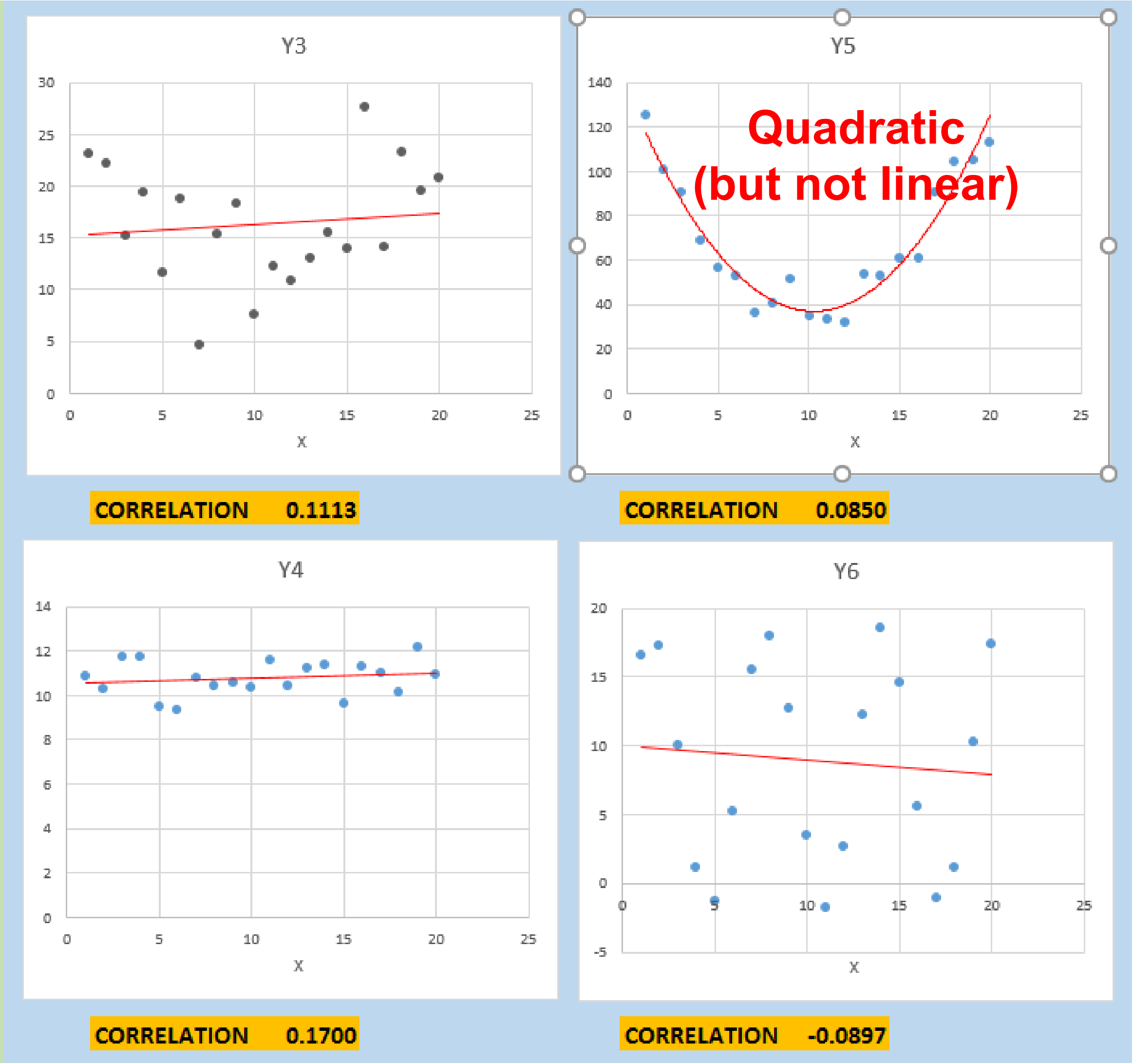
CORRELATION 0.9885

negative

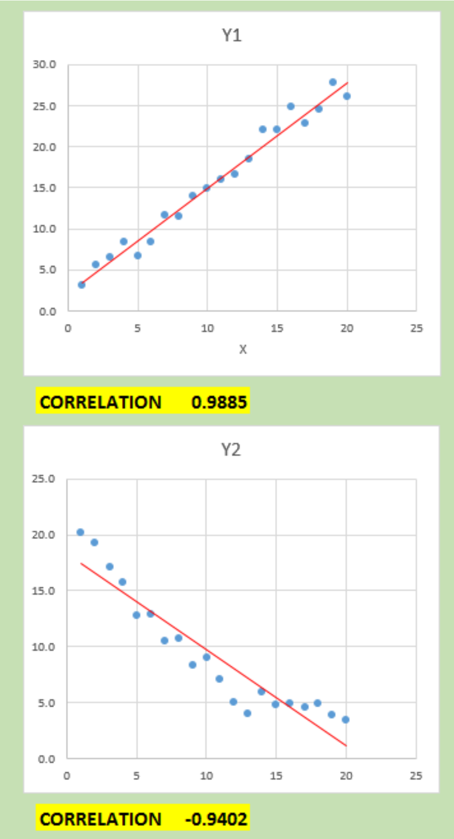


CORRELATION -0.9402

No linear relationship:



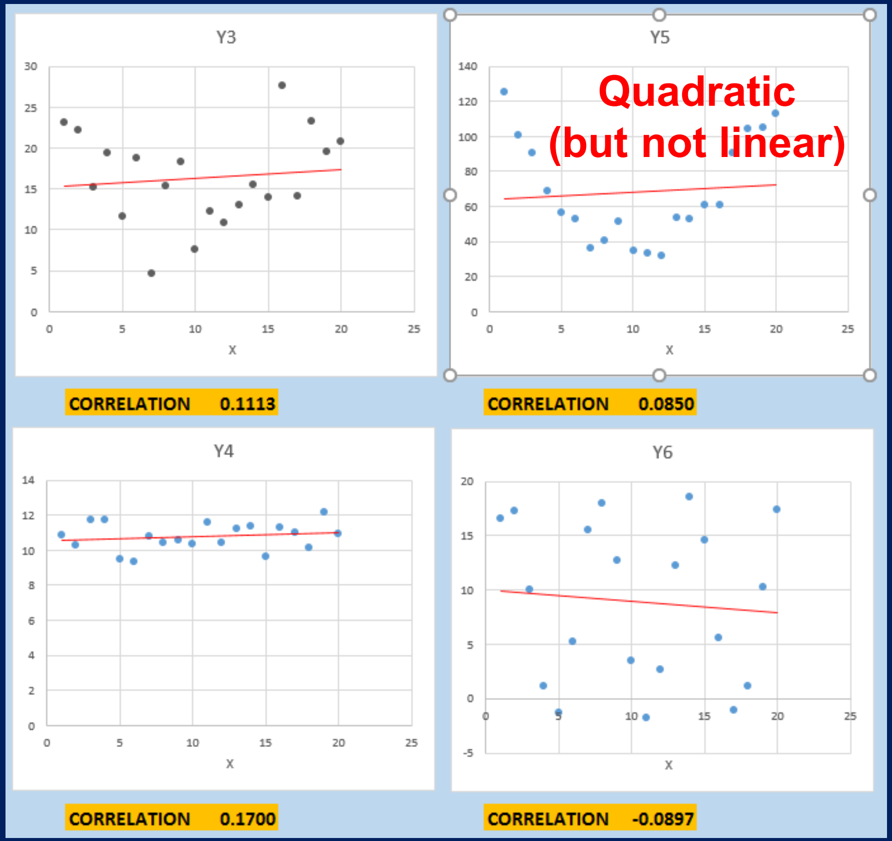
LINEAR relationship



Slope ≠ 0

X is a linear predictor of Y

NO LINEAR relationship

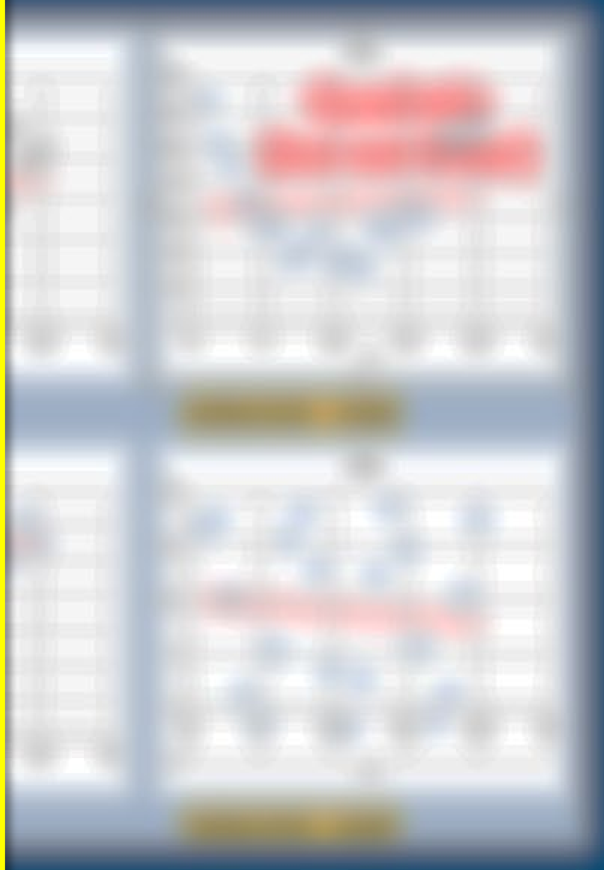


Slope = 0

X is **NOT** a **linear** predictor of Y

When we say “slope ≠ 0” or “slope=0”, we mean there is or there isn’t a relationship **in the STATISTICAL SENSE**.

In other words, this relationship should apply to the **entire POPULATION** of data, not just a sample.



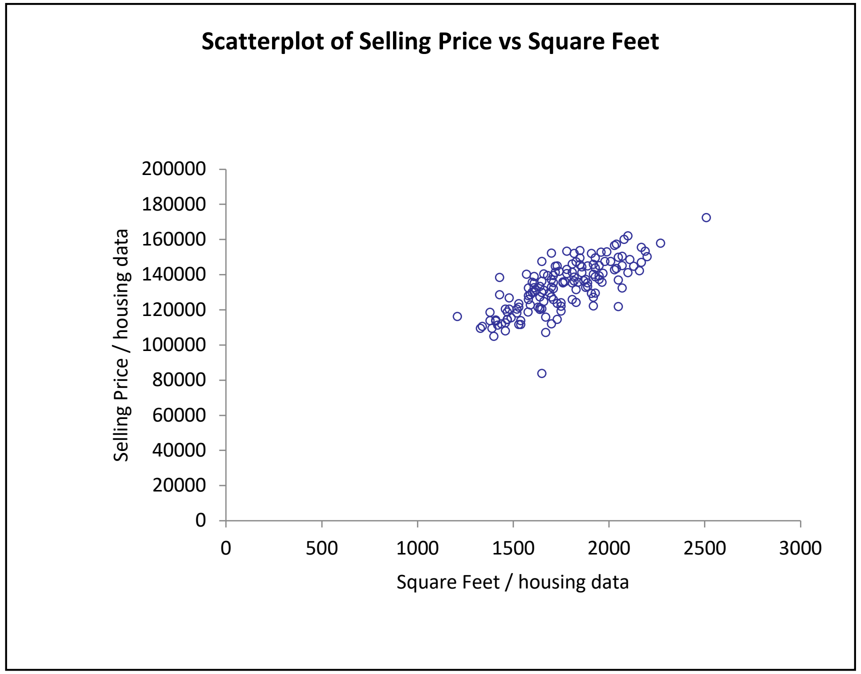
NULL hypothesis (H_0): **SLOPE = 0** **(X is not a good linear predictor of Y)**

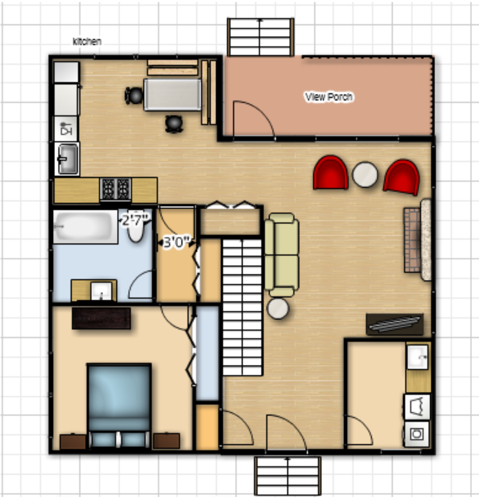
ALTERNATIVE hypothesis (H_A): **SLOPE ≠ 0** **(X is a good linear predictor of Y)**

- Here, when we say “SLOPE”, we mean the **true slope** for the entire **population**
- It’s a two-tailed test
- We want to reject the Ho
- To reject Ho, we want: t-statistic to be high and p-value to be low

How does the test work?

Recall Excel data **Housing data 148.xlsx**





$r = 0.749$

Simple linear regression:

$\hat{y} = 50527 + 46.991x$

$R^2 = 0.5618$

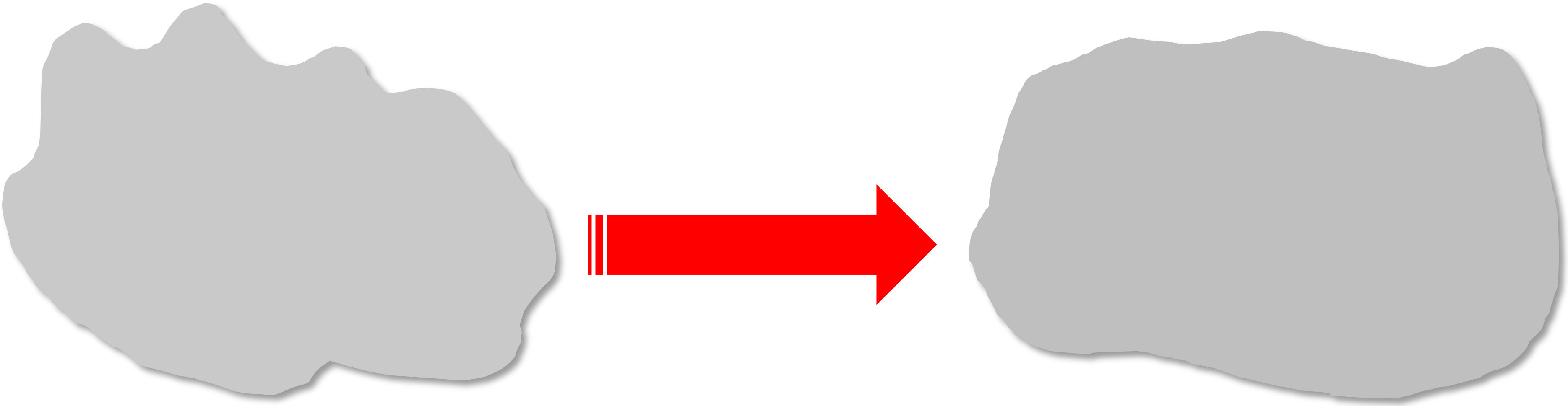
These coefficients were obtained from one **sample** of 148 homes

X 

square footage

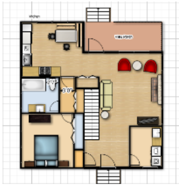
Y 

price



Population of **all** houses

Population of **all** houses



square footage

TRUE INTERCEPT for the population

TRUE SLOPE for the population

price

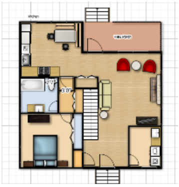
$\hat{Y}$ = **intercept** + **slope*X**

??? **???**

Population of **all** houses

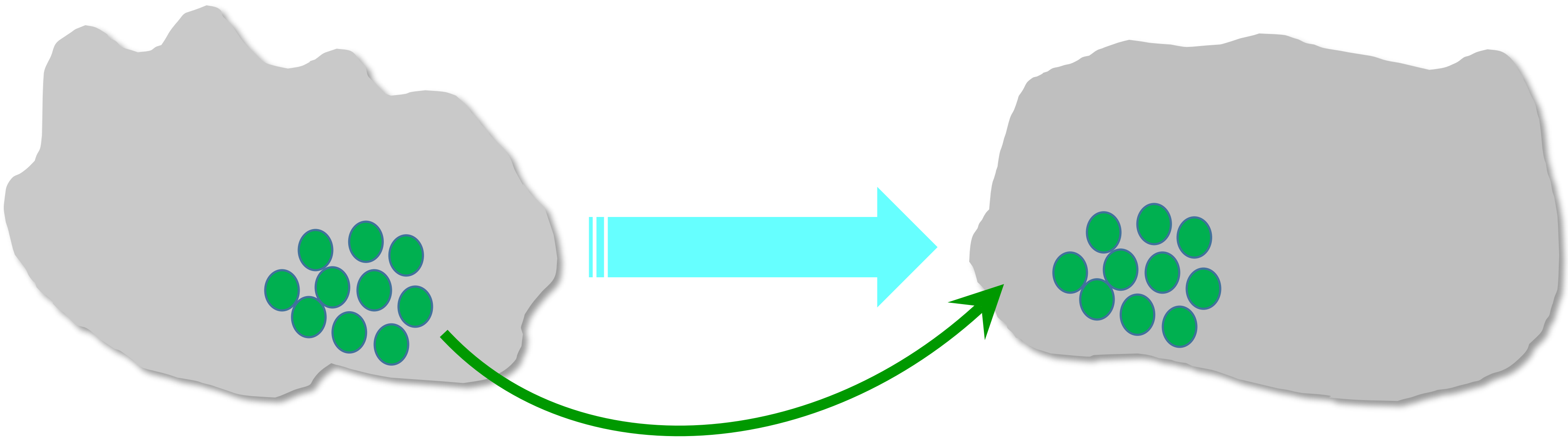
Population of **all** houses

But our data of 148 homes is just a **sample** from this population…



square footage

price



Our Excel data sample (148 homes):

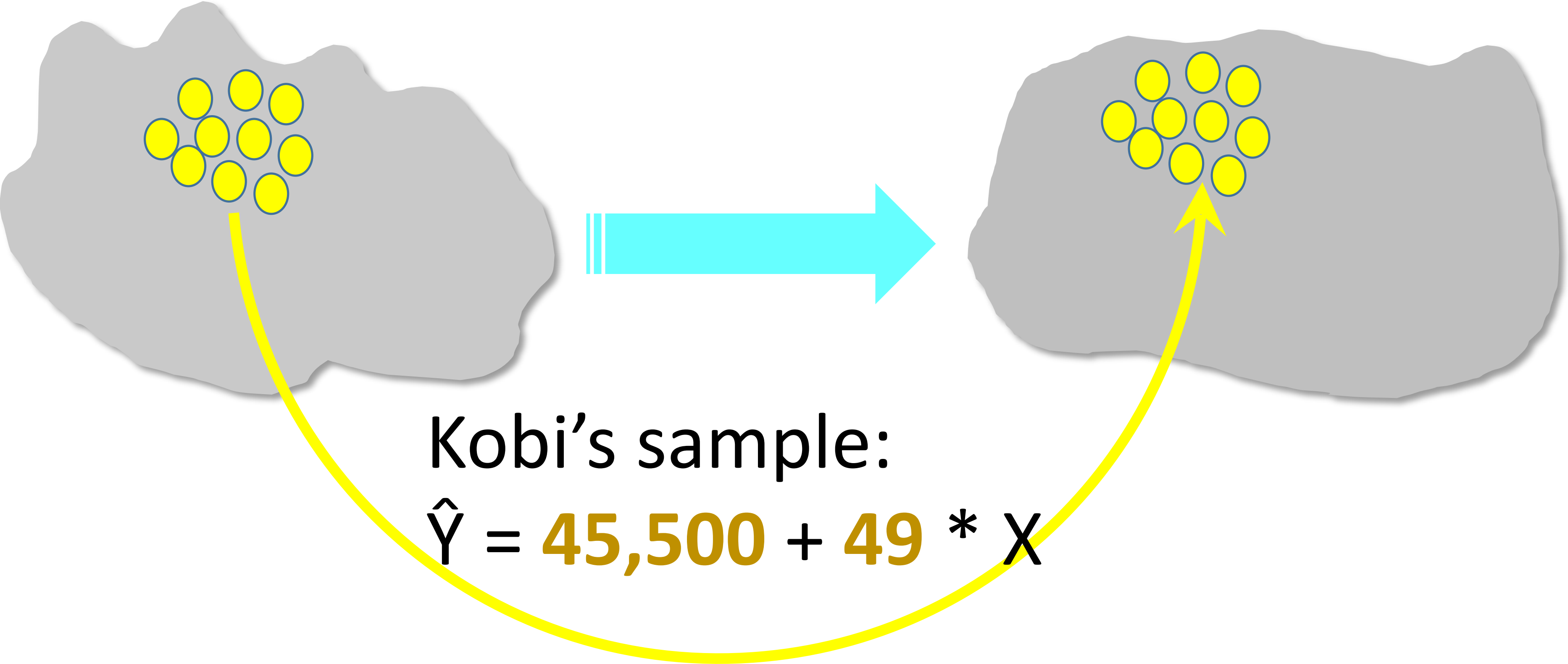
$\hat{Y} = \mathbf{50,527} + \mathbf{46.991}\, X$

What if we pick a different sample?

Would the coefficients remain the same?

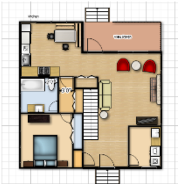
NO !

square footage **price**



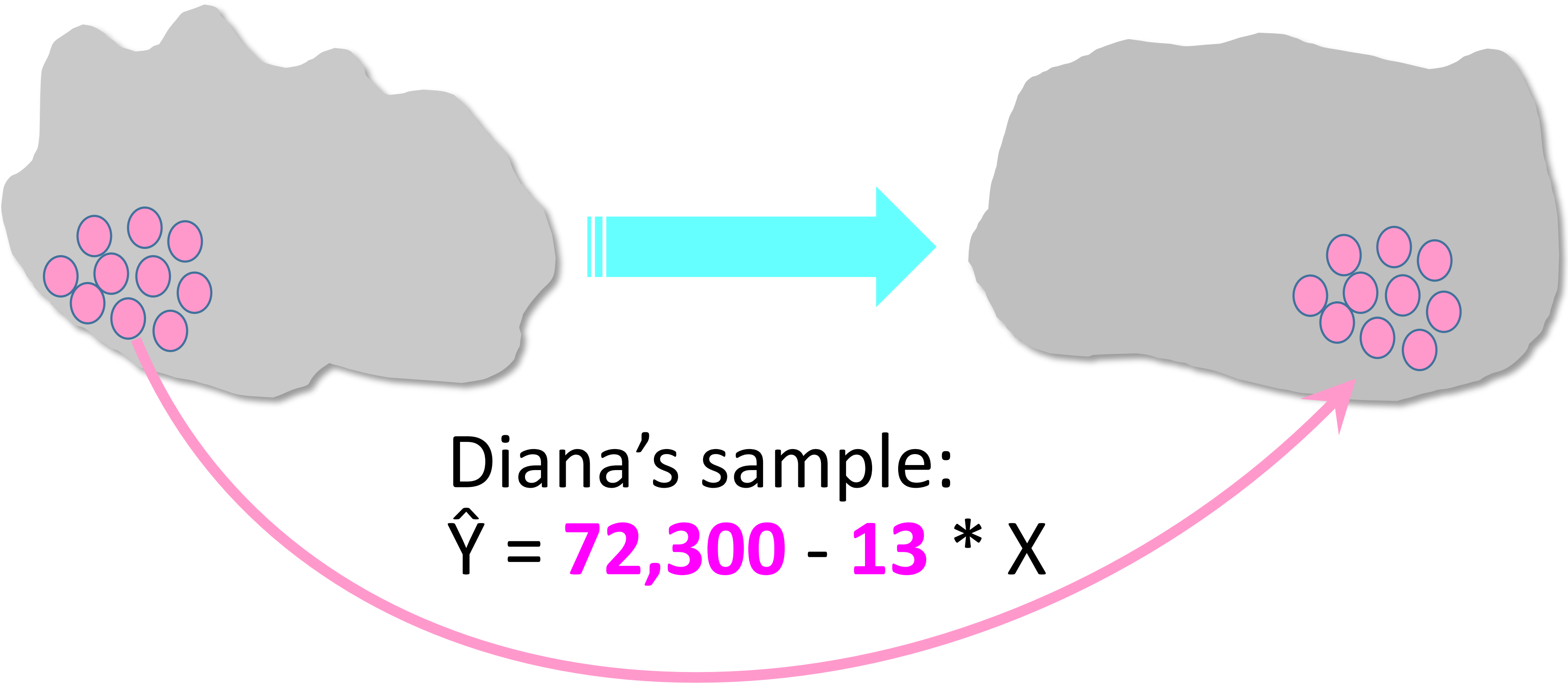
Kobi’s sample:

$\hat{Y}$ = **45,500** + **49** * X



square footage

price



Diana's sample:

$\hat{Y}$ = **72,300** - **13** * X

Depending on which samples of X and Y we pick, the intercept and slope coefficients would be different.

POPULATION

$\hat{Y} = \beta_0 + \beta_1 * X$

$\hat{Y} = ??? + ??? * X$

These are true (population) coefficients.

But they are unknown to us.

SAMPLE 1 — $\hat{Y} = b0 + b1 * X$ — $\hat{Y} = 45{,}500 + 49 * X$

SAMPLE 2 — $\hat{Y} = b0 + b1 * X$ — $\hat{Y} = 52{,}800 + 25 * X$

SAMPLE 3 — $\hat{Y} = b0 + b1 * X$ — $\hat{Y} = 72{,}300 - 13 * X$

Other samples — … — …

Overall, do we have evidence that the **true population slope ≠ 0** ?

In other words, do we have evidence that X is a good linear predictor of Y?

Reject Ho if *p*-value < α

Conclusion if ***p*-value < α**:

Yes, we **have** sufficient sample evidence that X is a good *linear* predictor of Y (or, that yes, there is a *linear* relationship between X and Y)

Conclusion if ***p*-value > α**:

No, we **don't have** sufficient sample evidence that X is a good *linear* predictor of Y (or, that no, we have no evidence of a *linear* relationship between X and Y)

Let’s look at our multiple regression example from a few weeks ago…

	A	B	C	D	E	F
1	House	Appraised Value	Selling Price	Square Feet	Bedrooms	Bathrooms
2	1	119,370	121,870	2050	4	5
3	2	148,930	150,250	2200	4	4
4	3	130,390	122,780	1590	3	3
5	4	135,700	144,350	1860	3	3
6	5	126,300	116,200	1210	2	3
7	6	137,080	139,490	1710	3	2
8	7	123,490	115,730	1670	3	3
9	8	150,830	140,590	1780	3	4
10	9	123,480	120,290	1520	4	4
11	10	132,050	147,250	1830	2	3
12	11	148,210	152,260	1700	3	3
13	12	139,530	144,800	1720	3	4
14	13	114,340	107,060	1670	3	4
15	14	140,040	147,470	1650	3	3
16	15	136,010	135,120	1610	2	1
17	16	140,930	140,240	1570	3	4
18	17	132,420	129,890	1650	4	5
19	18	118,300	121,140	1640	3	4
20	19	122,140	111,230	1420	2	3
21	20	149,820	145,140	2070	4	3
22	21	128,910	139,010	1610	2	3
23	22	134,610	129,340	1910	4	4
24	23	121,990	113,610	1410	2	2
25	24	150,500	141,050	1860	4	3
26	25	142,870	152,900	1990	4	3
27	26	155,550	157,790	2270	5	4
28	27	128,500	135,570	1965	4	4
29	28	143,360	151,990	1820	3	3
30	29	119,650	120,530	1650	3	3
31	30	122,570	118,640	1470	2	2
32	31	145,270	149,510	1850	4	3

Y X_1 X_2 X_3

Example

Housing Prices

Excel file **Housing Data 148.xlsx**

We want to understand how house selling price is explained by square footage, number of bedrooms, and number of baths.

Y = Selling prices

$\mathbf{X_1}$ = Square footage

$\mathbf{X_2}$ = Number of bedrooms

$\mathbf{X_3}$ = Number of bathrooms

	A	B	C	D	E	F
1	House	Appraised Value	Selling Price	Square Feet	Bedrooms	Bathrooms
2	1	119,370	121,870	2050	4	5
3	2	148,930	150,250	2200	4	4
4	3	130,390	122,780	1590	3	3
5	4	135,700	144,350	1860	3	3
6	5	126,300	116,200	1210	2	3
7	6	137,080	139,490	1710	3	2
8	7	123,490	115,730	1670	3	3
9	8	150,830	140,590	1780	3	4
10	9	123,480	120,290	1520	4	4
11	10	132,050	147,250	1830	2	3
12	11	148,210	152,260	1700	3	3
13	12	139,530	144,800	1720	3	4
14	13	114,340	107,060	1670	3	4
15	14	140,040	147,470	1650	3	3
16	15	136,010	135,120	1610	2	1
17	16	140,930	140,240	1570	3	4
18	17	132,420	129,890	1650	4	5
19	18	118,300	121,140	1640	3	4
20	19	122,140	111,230	1420	2	3
21	20	149,820	145,140	2070	4	3
22	21	128,910	139,010	1610	2	3
23	22	134,610	129,340	1910	4	4
24	23	121,990	113,610	1410	2	2
25	24	150,500	141,050	1860	4	3
26	25	142,870	152,900	1990	4	3
27	26	155,550	157,790	2270	5	4
28	27	128,500	135,570	1965	4	4
29	28	143,360	151,990	1820	3	3
30	29	119,650	120,530	1650	3	3
31	30	122,570	118,640	1470	2	2
32	31	145,270	149,510	1850	4	3

Y X_1 X_2 X_3

Predicted selling price ($) = 50,024 + 47.6 * Square Feet — 136.5 * #Bedrooms — 75.8 * #Baths

Multiple Regression for Selling Price Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.7496	0.5618	0.5527	9612.999346		
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
Explained	3	17062567906	5687522635	61.5468	< 0.0001	
Unexplained	144	13307004927	92409756.43			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	**50023.77969**	6951.207722	7.1964	< 0.0001	36284.19583	63763.36355
Square Feet	**47.64512751**	5.650129736	8.4326	< 0.0001	36.47722196	58.81303307
Bedrooms	**-136.4500702**	1780.582079	-0.0766	0.9390	-3655.904211	3383.004071
Bathrooms	**-75.79511324**	941.8675652	-0.0805	0.9360	-1937.467058	1785.876832

t-statistics:

The higher (in absolute value) the better

	A	B	C	D	E	F
1	House	Appraised Value	Selling Price	Square Feet	Bedrooms	Bathrooms
2	1	119,370	121,870	2050	4	5
3	2	148,930	150,250	2200	4	4
4	3	130,390	122,780	1590	3	3
5	4	135,700	144,350	1860	3	3
6	5	126,300	116,200	1210	2	3
7	6	137,080	139,490	1710	3	2
8	7	123,490	115,730	1670	3	3
9	8	150,830	140,590	1780	3	4
10	9	123,480	120,290	1520	4	4
11	10	132,050	147,250	1830	2	3
12	11	148,210	152,260	1700	3	3
13	12	139,530	144,800	1720	3	4
14	13	114,340	107,060	1670	3	4
15	14	140,040	147,470	1650	3	3
16	15	136,010	135,120	1610	2	1
17	16	140,930	140,240	1570	3	4
18	17	132,420	129,890	1650	4	5
19	18	118,300	121,140	1640	3	4
20	19	122,140	111,230	1420	2	3
21	20	149,820	145,140	2070	4	3
22	21	128,910	139,010	1610	2	3
23	22	134,610	129,340	1910	4	4
24	23	121,990	113,610	1410	2	2
25	24	150,500	141,050	1860	4	3
26	25	142,870	152,900	1990	4	3
27	26	155,550	157,790	2270	5	4
28	27	128,500	135,570	1965	4	4
29	28	143,360	151,990	1820	3	3
30	29	119,650	120,530	1650	3	3
31	30	122,570	118,640	1470	2	2
32	31	145,270	149,510	1850	4	3

Y X_1 X_2 X_3

Predicted selling price ($) = 50,024 + 47.6 * Square Feet — 136.5 * #Bedrooms — 75.8 * #Baths

Multiple Regression for Selling Price *Summary*	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.7496	0.5618	0.5527	9612.999346		
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
Explained	3	17062567906	5687522635	61.5468	< 0.0001	
Unexplained	144	13307004927	92409756.43			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	**50023.77969**	6951.207722	7.1964	< 0.0001	36284.19583	63763.36355
Square Feet	**47.64512751**	5.650129736	8.4326	< 0.0001	36.47722196	58.81303307
Bedrooms	**-136.4500702**	1780.582079	-0.0766	0.9390	-3655.904211	3383.004071
Bathrooms	**-75.79511324**	941.8675652	-0.0805	0.9360	-1937.467058	1785.876832

P-values:

The closer to 0 the better

(should be **< α** , e.g., 0.05)

	A	B	C	D	E	F
1	House	Appraised Value	Selling Price	Square Feet	Bedrooms	Bathrooms
2	1	119,370	121,870	2050	4	5
3	2	148,930	150,250	2200	4	4
4	3	130,390	122,780	1590	3	3
5	4	135,700	144,350	1860	3	3
6	5	126,300	116,200	1210	2	3
7	6	137,080	139,490	1710	3	2
8	7	123,490	115,730	1670	3	3
9	8	150,830	140,590	1780	3	4
10	9	123,480	120,290	1520	4	4
11	10	132,050	147,250	1830	2	3
12	11	148,210	152,260	1700	3	3
13	12	139,530	144,800	1720	3	4
14	13	114,340	107,060	1670	3	4
15	14	140,040	147,470	1650	3	3
16	15	136,010	135,120	1610	2	1
17	16	140,930	140,240	1570	3	4
18	17	132,420	129,890	1650	4	5
19	18	118,300	121,140	1640	3	4
20	19	122,140	111,230	1420	2	3
21	20	149,820	145,140	2070	4	3
22	21	128,910	139,010	1610	2	3
23	22	134,610	129,340	1910	4	4
24	23	121,990	113,610	1410	2	2
25	24	150,500	141,050	1860	4	3
26	25	142,870	152,900	1990	4	3
27	26	155,550	157,790	2270	5	4
28	27	128,500	135,570	1965	4	4
29	28	143,360	151,990	1820	3	3
30	29	119,650	120,530	1650	3	3
31	30	122,570	118,640	1470	2	2
32	31	145,270	149,510	1850	4	3

Y X_1 X_2 X_3

Predicted selling price ($) = 50,024 + 47.6 * Square Feet — 136.5 * #Bedrooms — 75.8 * #Baths

Multiple Regression for Selling Price Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.7496	0.5618	0.5527	9612.999346		
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
Explained	3	17062567906	5687522635	61.5468	< 0.0001	
Unexplained	144	13307004927	92409756.43			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	**50023.77969**	6951.207722	7.1964	< 0.0001	36284.19583	63763.36355
Square Feet	**47.64512751**	5.650129736	8.4326	< 0.0001	36.47722196	58.81303307
Bedrooms	**-136.4500702**	1780.582079	-0.0766	0.9390	-3655.904211	3383.004071
Bathrooms	**-75.79511324**	941.8675652	-0.0805	0.9360	-1937.467058	1785.876832

Which of these 3 variables are good *linear* predictors of selling price?

We can extract the same information from the CONFIDENCE INTERVALS for the true (population) slope.

Multiple Regression for Selling Price Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.7496	0.5618	0.5527	9612.999346		
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
Explained	3	17062567906	5687522635	61.5468	< 0.0001	
Unexplained	144	13307004927	92409756.43			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	**50023.77969**	6951.207722	7.1964	< 0.0001	36284.19583	63763.36355
Square Feet	**47.64512751**	5.650129736	8.4326	< 0.0001	36.47722196	58.81303307
Bedrooms	**-136.4500702**	1780.582079	-0.0766	0.9390	-3655.904211	3383.004071
Bathrooms	**-75.79511324**	941.8675652	-0.0805	0.9360	-1937.467058	1785.876832

Slope (b) estimated from this sample

Multiple Regression for Selling Price Summary	Multiple R	R-Square	Adjusted R-Square	StErr of Estimate		
	0.7496	0.5618	0.5527	9612.999346		
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F-Ratio	p-Value	
Explained	3	17062567906	5687522635	61.5468	< 0.0001	
Unexplained	144	13307004927	92409756.43			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	**50023.77969**	6951.207722	7.1964	< 0.0001	36284.19583	63763.36355
Square Feet	**47.64512751**	5.650129736	8.4326	< 0.0001	36.47722196	58.81303307
Bedrooms	**-136.4500702**	1780.582079	-0.0766	0.9390	-3655.904211	3383.004071
Bathrooms	**-75.79511324**	941.8675652	-0.0805	0.9360	-1937.467058	1785.876832

True slope lope (β) for the entire population

The interval for **Square Feet coefficient** means:

"We are 95% confident that the **true (population) coefficient** of Square Feet is between 36.48 and 58.81."

This interval does *not* contain zero. This means that we have evidence that the true slope is different from zero.

Application to MODEL SELECTION.

Stepwise regression

STEPWISE REGRESSION:
Include X variables one at a time in the following order: SqFt, BR, BA

Step 1: Y=price, X=SqFt

				R-Square	Adjusted R-square	
				0.5618	0.5587	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50526.84317	6076.382246	8.315283852	< 0.0001	38517.81152	62535.87482
Square Feet	46.99097199	3.435004385	13.6800326	< 0.0001	40.202216	53.77972798

Step 2: Y=price, X1=SqFt, X2=#BR

Drop #BR

					Adjusted R-square	
					0.555767	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50064.47532	6908.996023	7.246273576	< 0.0001	36409.12406	63719.82658
Square Feet	47.62327294	5.624231373	8.467516677	< 0.0001	36.50720706	58.73933882
Bedrooms	-213.1917728	1498.522321	-0.142268	0.8871	-3174.960544	2748.576998

Step 3: Y=price, X1=SqFt, X2=#BA

Drop #BA

					Adjusted R-square	
					0.5557692	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50214.07161	6470.24661	7.760766264	< 0.0001	37425.89082	63002.25239
Square Feet	47.36766076	4.322627246	10.95807204	< 0.0001	38.82416281	55.91115872
Bathrooms	-114.4513591	792.665853	-0.144387901	0.8854	-1681.123364	1452.220646

FINAL MODEL:

				R-Square	Adjusted R-square	
				0.5618	0.5587	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50526.84317	6076.382246	8.315283852	< 0.0001	38517.81152	62535.87482
Square Feet	46.99097199	3.435004385	13.6800326	< 0.0001	40.202216	53.77972798

What do the p-values mean? Assume α=0.05

***P*-value=0** < α . Reject Ho. Yes, we have evidence of a **linear** relationship between SqFt and Price.

				R-Square	Adjusted R-square	
				0.5618	0.5587	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50526.84317	6076.382246	8.315283852	< 0.0001	38517.81152	62535.87482
Square Feet	46.99097199	3.435004385	13.6800326	< 0.0001	40.202216	53.77972798

***P*-value = 0.8871** > α. Don't reject Ho. No evidence of a **linear** relationship between #BR and Price. (Nonlinear?? Maybe…)

					Adjusted R-square	
					0.555767	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50064.47532	6908.996023	7.246273576	< 0.0001	36409.12406	63719.82658
Square Feet	47.62327294	5.624231373	8.467516677	< 0.0001	36.50720706	58.73933882
Bedrooms	-213.1917728	1498.522321	-0.142268	0.8871	-3174.960544	2748.576998

***P*-value = 0.8854** > α. Don't reject Ho. No evidence of a **linear** relationship between #BA and Price. (Nonlinear?? Maybe…)

					Adjusted R-square	
					0.5557692	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Upper
Constant	50214.07161	6470.24661	7.760766264	< 0.0001	37425.89082	63002.25239
Square Feet	47.36766076	4.322627246	10.95807204	< 0.0001	38.82416281	55.91115872
Bathrooms	-114.4513591	792.665853	-0.144387901	0.8854	-1681.123364	1452.220646

Backward elimination

BACKWARD ELIMINATION:
Include all X variables, then remove one at a time in the order: #BA, #BR

Step 1: Y=price X1=SqFt
X2=#BR
X3=#BA

Multiple Regression for Selling Price	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
Summary	0.7496	0.5618	0.552702494	9612.999346	0	0

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	50023.77969	6951.207722	7.196415599	< 0.0001	36284.19583	63763.36355
Square Feet	47.64512751	5.650129736	8.432572301	< 0.0001	36.47722196	58.81303307
Bedrooms	-136.4500702	1780.582079	-0.076632283	0.9390	-3655.904211	3383.004071
Bathrooms	-75.79511324	941.8675652	-0.080473217	0.9360	-1937.467058	1785.876832

Step 2: Remove #BA because p-value is high
Y=price X1=SqFt
X2=#BR

Multiple Regression for Selling Price	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
Summary	0.7495	0.5618	0.555767327	9580.009131	0	0

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	50064.47532	6908.996023	7.246273576	< 0.0001	36409.12406	63719.82658
Square Feet	47.62327294	5.624231373	8.467516677	< 0.0001	36.50720706	58.73933882
Bedrooms	-213.1917728	1498.522321	-0.142268	0.8871	-3174.960544	2748.576998

Step 3: Remove #BR because p-value is high
Y=price X1=SqFt
⇨ Final model

Multiple Regression for Selling Price	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
Summary	0.7495	0.5618	0.5587484	9547.810816	0	0

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	50526.84317	6076.382246	8.315283852	< 0.0001	38517.81152	62535.87482
Square Feet	46.99097199	3.435004385	13.6800326	< 0.0001	40.202216	53.77972798

use p-values

BACKWARD ELIMINATION:

Include all X variables, then remove one at a time ~~in the order: #BA, #BR~~

<u>Step 1:</u> Y=price X1=SqFt
X2=#BR
X3=#BA

Multiple Regression for Selling Price *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.7496	0.5618	0.552702494	9612.999346	0	0
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	50023.77969	6951.207722	7.196415599	< 0.0001	36284.19583	63763.36355
Square Feet	47.64512751	5.650129736	8.432572301	< 0.0001	36.47722196	58.81303307
Bedrooms	-136.4500702	1780.582079	-0.076632283	0.9390	-3655.904211	3383.004071
Bathrooms	-75.79511324	941.8675652	-0.080473217	0.9360	-1937.467058	1785.876832

<u>Step 2:</u> Remove #BR
because p-value is the highest

<u>Step 3:</u> Remove #BA
because p-value is the second highest

⇨ Final model:
Y=price X1=SqFt

Multiple Regression for Selling Price *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.7495	0.5618	0.5587484	9547.810816	0	0
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper
Constant	50526.84317	6076.382246	8.315283852	< 0.0001	38517.81152	62535.87482
Square Feet	46.99097199	3.435004385	13.6800326	< 0.0001	40.202216	53.77972798

Hypothesis testing in regressions